DOCUMENT RESUME

ED 430 035 TM 029 779

AUTHOR Sanford, Eleanor E.; Fabrizio, Louis M.
TITLE Results from the North Carolina--NAEP Comparison and What They Mean to the End-of-Grade Testing Program.
PUB DATE 1999-04-00
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Educational Assessment; *Grade 8; *Judges; Junior High Schools; *Mathematics Tests; National Surveys; *Test Results; *Testing Programs
IDENTIFIERS Experts; *National Assessment of Educational Progress; *North Carolina End of Grade Testing Program; State Competency Tests

ABSTRACT

The North Carolina End-of-Grade (NCEOG) Test of Mathematics, Grade 8 was compared to the National Assessment of Educational Progress (NAEP)--Grade 8 along three dimensions: technical, content, and cognitive. An expert panel of six judges (mathematics educators) examined the supporting documents, the content frameworks, the items, and actual test forms to determine the level of congruence between the two assessments. Although differences were observed between the two assessments on each of the three dimensions, they were not enough to explain the differences in student performance on the assessments. The most important result of the study was a clearer understanding of the NCEOG, and following that an improved understanding of the NAEP. Further research to investigate the differences between the two assessments is outlined. (Contains 4 tables and 14 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 035

Results from the North Carolina–NAEP Comparison and What They Mean to the End-of-Grade Testing Program

Eleanor E. Sanford
MetaMetrics, Inc.

Louis M. Fabrizio
North Carolina Department of Public Instruction

April 1999

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Eleanor Sanford
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada

TM029779

BEST COPY AVAILABLE

Results from the North Carolina–NAEP Comparison and What They Mean to the End-of-Grade Testing Program

The North Carolina End-of-Grade Test of Mathematics, Grade 8 (NCEOG) was compared to the National Assessment of Educational Progress, Mathematics—Grade 8 (NAEP) along three dimensions—technical, content, and cognitive. Expert panels of judges examined the supporting documents, the content frameworks, the items, and actual test forms to determine the level of congruence between the two assessments. While differences were observed between the two assessments on each of the three dimensions, those differences were not enough to explain the differences in student performance on the assessments.

Introduction

In 1996 Mark Musick compared state-level results from the National Assessment of Educational Progress (NAEP) with results reported by states on state assessments and found wide variations in the proportion of students reported as "proficient" (Musick, 1996; Archer, 1997). He concluded that these variations were probably not due to "what states believe should be taught but in how much they expect students to learn" (Musick, 1996, p. 2). Without a standard indicator, or reference point, there is no real way to understand the results. Musick suggests that "the standards in many cases are so different that state leaders and those in charge of the National Assessment [NAEP] need to be around the same table seeking to understand the differences and whether changes are needed" (p. 3).

In part because of the disparities between the results of the NAEP and state assessments, a group was convened to identify options to improve the design of NAEP. As part of the Redesign Policy, adopted in August 1996, the National Assessment Governing Board outlined a number of goals and objectives for guiding changes in the National Assessment of Educational Progress. One goal related to state linking: To help states and others link their assessments with the National Assessment and use National Assessment data to improve educational performance. The policy also provides further specifications regarding the linking issue: The National Assessment shall develop policies, practices, and procedures that assist states, school districts, and others who want to do so at their own cost to link their test results to the National Assessment.

In North Carolina there has been a move to higher standards of proficiency in reading and mathematics during the 1990s. Much of this emphasis has resulted from North Carolina's improvement on the NAEP tests. When the 1996 results were released, "North Carolina fourth graders bested the national average for the first time, posting a gain since 1992 that tripled the national gain. The state's fourth graders tied with Texas for showing the highest gain in the nation, 11 points, since the last time the test was given in 1992. North Carolina eighth graders were three points below the national average, but above the Southeast average for math. Eighth-grade scores did show a nine-point gain from 1992 and were up 17 points from 1990. The 17-point gain since 1990 was the highest in the nation" (NCDPI, 1999). But [and this is a big "but"], North Carolina is still not at the level of results reported from the National Assessment of Educational Progress (NAEP) in terms of the students scoring at or above "proficient" level, however proficient is defined.

There are many possible explanations for the observed differences between the reported results of student performance on the North Carolina and NAEP assessments. One explanation has to do with the content frameworks of the assessments—they are two different assessments each based on a specified content framework. Another explanation has to do with how the results are reported; both assessments use the same general names for the performance standard levels but the descriptions are different. A third explanation has to do with the methods used to set the performance standards; different standard-setting methods typically produce different standards.

In February 1997 the North Carolina Department of Public Instruction (NCDPI) staff met with staff and members of the National Assessment Governing Board (NAGB). The purpose of the meeting was to discuss the apparent discrepancies in the percentages of students reported as scoring at the "proficient" level on the North Carolina assessments and the National Assessment of Educational Progress (NAEP) (NCDPI, 1997). Much of the meeting was spent in understanding the characteristics of the two assessments—how they compared and contrasted in terms of standard-setting procedures and the use of the results. The remainder of the meeting was spent discussing ways that North Carolina and NAGB could work together to examine the differences in standard setting methods employed with the two assessments. The following activities were recommended to better understand the differences:

- Review the content standards—the NC and NAEP content and test specifications should be examined for alignment.
- Review the performance standards—the NC and NAEP assessments should be incorporated into one standard setting session using the modified-Angoff procedure used with the NAEP assessment.
- Link the North Carolina assessments with other NAEP assessments.

Based on the meeting in Raleigh, the Governing Board offered to fund a study that would examine the relationship between the North Carolina grade eight mathematics assessment and that used by NAEP. This study was conducted by the staff of the Learning Research and Development Center of the University of Pittsburgh for the National Assessment Governing Board. The study was designed to examine the first possible explanation for the observed differences in performance between the two assessments—the content frameworks. The content and test specifications of the two mathematics assessments at grade 8 were to be examined to see if those differences could account for the differences in performance on the two assessments. The primary purposes of this study were:

- to examine the relationship between the framework, specifications, and test items used in the North Carolina End-of-Grade Test of Mathematics in grade eight and the NAEP grade eight math assessment; and
- to develop a model process that could be used by states, school districts, and others to compare their frameworks and assessments to NAEP.

Method

Congruence Dimensions

Three dimensions, or perspectives, common to the state test and NAEP were identified as relevant to this study. The *technical* dimension involves components such as the number and type of items, the time allotted for administering the test, the difficulty of the items, etc. The *content* dimension has to do with the particular content topics (e.g., for mathematics—geometry,

measurement, algebra) included. The *cognitive* dimension involves the extent to which a test engages students in various cognitive processes, including problem solving, reasoning, or the recall of facts and definitions. [For a complete discussion of the three dimensions used to compare the assessments and the activities used to assess the dimensions, refer to "Design Features for the Content Analysis of a State Assessment and NAEP" by Kenney and Silver (1999).]

Expert Panel

The panel of experts consisted of six mathematics education professionals (e.g., mathematics teachers, college/university mathematics educators, and mathematics curriculum specialists). The composition of the panel reflected distributed expertise that spanned the state test, NAEP, and middle school mathematics.

Of the six members, two members were selected on the basis of their familiarity with the state assessment; that is, they served the capacity that ensured knowledge of the state's testing program (e.g., serving on the mathematics framework development committee, writing test items, providing professional development for mathematics teachers on the state assessment program). The two "state" panelists could provide information related to the state test, should the need arise.

Another pair of panelists were selected on the basis of their knowledge of the NAEP mathematics assessment, and in particular the NAEP grade 8 test. For example, these panelists had served on committees that developed the NAEP mathematics framework and items, or they knew about NAEP through their involvement with other NAEP-related projects. The two "NAEP" panelists could provide information related to the NAEP test, should the need arise.

The final two panelists were selected for their expertise about and experiences with middle school mathematics education and for their lack of specialized knowledge about either the state assessment or about NAEP. The role of these panelists within the group was one of neutrality with respect to the tests to be examined; that is, this pair of "neutral" panelists had no vested interest in either test.

North Carolina End-of-Grade Tests

The North Carolina end-of-grade testing program was established in response to legislation passed by the 1989 North Carolina General Assembly. The tests assess reading comprehension and mathematics and were developed for two purposes:

- to provide accurate measurement of individual student skills and knowledge specified in the North Carolina *Standard Course of Study*, and
- to provide accurate measurement of the knowledge and skills attained by groups of students for school, school system, and state accountability (NCDPI, 1996).

All students in grades 3 through 8 are administered both assessments. The assessments are presented in multiple-choice format and the results are used for school-level accountability (grades 3 through 8) and student-level accountability (grade 8 as a competency screening for high school graduation). For school-level accountability, additional tests were developed for administration at the beginning of grade 3 (Fall 1996) and at the end of grade 10 (April 1998).

Performance standards provide a common meaning of test scores throughout the state concerning what is expected at various levels of competence. Performance standards were developed for the end-of-grade tests using the contrasting groups method. During the field test (May 1992), teachers were asked to categorize each student participating in the field test into one of four proficiency levels. Teachers were asked to base their judgements on their first-hand knowledge of the student's level of achievement during the school year in various domains assessed outside of the testing situation. Teachers are able to make informed judgements about students' achievement because the teachers have observed the breadth and depth of the work each student has accomplished during the school year. The four achievement levels are (emphasis added):

- Level I: Students performing at this level *do not have sufficient mastery* of knowledge and skills in this subject area to be successful at the next grade level.
- Level II: Students performing at this level *demonstrate inconsistent mastery* of knowledge and skills that are fundamental in this subject area and that are minimally sufficient to be successful at the next grade level.
- Level III: Students performing at this level *consistently demonstrate mastery* of grade level subject matter and skills and are *well-prepared* for the next grade level.
- Level IV: Student's performing at this level consistently *perform in a superior manner* clearly beyond that required to be proficient at grade level work.

The percentage of students categorized into each achievement level was applied to the distribution of scores when the tests were administered statewide for the first time (May 1993) and a range of scores was established for each achievement level at each grade. The range of scores associated with each achievement level have not been modified.

North Carolina End-of-Grade Test of Mathematics, Grade 8. The North Carolina End-of-Grade Test of Mathematics (NCEOG) consists of two parts: mathematics computation and mathematics applications. At the student level, the two parts of the test are combined to produce one mathematics score. The score is reported on a mathematics developmental scale (98-226) and percentiles were established based on the administration of the test in May 1993.

The mathematics computation part of the test (8 items) assesses a student's ability to do routine computations without a calculator. These items include symbolic computation skills and application skills such as estimation and word problems involving percents (tax, tip, sale price, etc.). The mathematics applications part of the test (72 items) assesses a student's ability to apply mathematical principles, solve problems, and explain mathematical processes. Problems are typically posed as real situations that students at the grade level may have encountered. Students are allowed to use calculators, rulers, and protractors on this part of the test.

The items for the itempool and the test were specified by goal and objective from the North Carolina *Standard Course of Study* (Mathematics, Grade 8) adopted by the North Carolina State Board of Education in June 1989 (NCDPI, 1989). Table 1 shows the content specifications for each part of the test and the test overall by curricular strand. Within goals, objectives were not weighted equally in the test specifications. Each objective was examined by the NCDPI mathematics curriculum specialists and weighted appropriately.

Table 1. Test Specifications for the North Carolina End-of-Grade Test of Mathematics, Grade 8.

Goal/ Strand	Description	Percent of Items (Applications/Computation)
1	*Numeration*: The learner will demonstrate an understanding and use of real numbers.	13.5 (11 and 2.5)
2	*Geometry*: The learner will demonstrate an understanding and use of properties and relationships of geometry.	10
3	*Patterns/Pre-Algebra*: The learner will demonstrate an understanding of pre-algebra.	17.5 (15 and 2.5)
4	*Measurement*: The learner will demonstrate an understanding and use of measurement.	10
5	*Problem Solving*: The learner will solve problems and reason mathematically.	15
6	*Statistics*: The learner will demonstrate an understanding and use of probability and statistics.	12
7	*Computation*: The learner will compute with real numbers.	21 (16 and 5)

In addition to specifying the content of each item during item development, the difficulty level and thinking skill for each item to be developed were also specified. Difficulty level describes how hard the item is. Items were specified to be easy, medium, or hard. This specification ensured that the item writers developed a range of items to assess each curricular objective independent of the difficulty of the specific content to be assessed. Thinking skill level describes the cognitive skills that a student must employ to solve the problem. The thinking skill framework used with the North Carolina End-of-Grade Tests is from *Dimensions of Thinking* by Robert J. Marzano and others (1988). This framework consists of 21 core thinking skills; a thinking skill is a relatively specific cognitive operation that can be considered a "building block" of thinking. These 21 core thinking skills can be organized into 7 broad skills—knowledge, organizing, applying, analyzing, generating, integrating, and evaluating.

Items on each final test form were not selected on the basis of difficulty level or thinking skill. This item-level information was initially specified for item development to ensure that the item pool contained a broad range of difficulty level and thinking skills to be employed to solve the problems and that items required more than rote learning. Items for the final test forms were selected on the basis of the curricular and psychometric characteristics of the items.

National Assessment of Educational Progress

The National Assessment of Educational Progress (NAEP) is a congressionally-mandated survey of achievement of the nation's students in grades 4, 8, and 12. NAEP assessments are administered every two years, with a specific subject such as reading, writing, mathematics, or science administrated once every four years. In 1990, NAEP began a voluntary state-by-state assessment program which allows states to compare their achievement with that of other states and the nation as a whole for grade 8. NAEP uses a representative probability sample based on students within schools within geographic areas. NAEP is generally perceived as a low-stakes assessment.

The NAEP mathematics test is aligned with the National Council of Teachers of Mathematics standards. The assessment is organized according to three mathematical abilities—conceptual understanding, procedural knowledge, and problem solving—and five content strands—

numbers and operations; measurement; geometry; data analysis, statistics, and probability; and algebra and functions. NAEP is not tied directly to any curriculum framework, but instead is a broad-based assessment of topics in the mathematics curriculum at grades 4, 8, and 12. The assessment is presented in both multiple-choice and constructed response formats.

NAEP results are reported for the nation and demographic subgroups and for states that voluntarily participate in the state-level assessment. Performance standards (achievement levels) were developed using the modified-Angoff method and are reported in terms of what students "should be able" to do.

- Basic: Eighth-grade students performing at the **basic level** *should exhibit evidence of conceptual and procedural understanding in the five NAEP content strands.* This level of performance signifies an understanding of arithmetic operations—including estimation—on whole numbers, decimals, fractions, and percents.
- Proficient: Eighth-grade students performing at the **proficient level** *should apply mathematical concepts and procedures consistently to complex problems in the five NAEP content strands.*
- Eighth-grade students at the **advanced level** *should be able to reach beyond the recognition, identification, and application of mathematical rules in order to generalize and synthesize concepts and principles in the five NAEP content strands.*

[For a more extensive overview of the National Assessment of Educational Progress and specifically the grade 8 mathematics assessment, refer to *Mathematics Framework for the 1996 National Assessment of Educational Progress* published by The College Board (1994).]

Results

The process for evaluating the congruence between the North Carolina End-of-Grade Test of Mathematics, Grade 8 (NCEOG) and the National Assessment of Educational Progress, Mathematics—Grade 8 (NAEP) involved a five-phase process and a variety of activities. Two of the five phases (Phases II and IV) consisted of the activities occurring during the two-day meetings of the expert panel. The other three phases consisted of the collection and review of documents related to the assessments, the preparation of materials for the congruence activities, the analysis of data generated during the activities, and the production of summaries of the activities and the meetings.

Technical Dimension

The first phase of the study was designed to compare the technical characteristics of the NCEOG and the NAEP assessments. The analysis of the NCEOG and the NAEP along the technical dimension was compiled by the LRDC project staff based on technical documents related to each assessment. The following documents were used:

- curricular frameworks—the *North Carolina Standard Course of Study for Mathematics* (NCDPI, 1989) and the 1996 NAEP mathematics framework document (The College Board, 1994).
- technical reports—the *North Carolina End-of-Grade Tests, Technical Report #1* (NCDPI, 1996).

Technical information was also obtained from presentations made during Phase II of the study. A representative from the NCDPI and a member of the LRDC staff presented information about each respective test. Each presentation consisted of an overview of the purpose of the test and its important technical characteristics.

The results of this analysis were shared with the panelists and other participants as needed during the meetings. Table 2 contains the results of this analysis.

Table 2. Comparison of the NCEOG and NAEP according to selected technical characteristics.

Technical Characteristic	NCEOG	NAEP
Item Format	Multiple-Choice	Multiple-Choice: 55% Constructed Response: 45%
Number of Items	80 (8 computation/72 applications)	Varies: Each student takes 3 blocks of items; number of items per block varies (10 to 20)
Distribution of Items by Content Area	*Measurement*: 10% *Geometry*: 10% *Probability & Statistics*: 12% *Patterns/Pre-Algebra*: 17.5% *Numeration*: 13.5% *Computation*: 21% *Problem-Solving*: 15%	*Measurement*: 15% *Geometry/Spatial Sense*: 20% *Data, Statistics, Probability*: 15% *Algebra and Functions*: 25% *Number Sense, Properties, & Operations*: 25%
Administration Time	97 minutes (12 computation/87 applications)	45 minutes (15 per block/3 blocks)
Items Administered at Multiple Grade Levels	No	Yes: subset administered at grades 4/8, grades 8/12, and grades 4/8/12

During their deliberation, the panelists and other participants became cognizant of the differences in the technical characteristics of the two tests, but they were not asked to make a congruence judgement based on the technical dimension. By design, these judgements were withheld until the panelists had the opportunity to view the tests along the content and cognitive dimensions.

Content Dimension

The second phase of the process involved examining the content characteristics of the NCEOG and the NAEP assessments. In particular, the content characteristics involved what was assessed on the test; that is, the mathematics topics and the coverage of each of the topics on the test itself. The relationship between the assessments along the content dimension was investigated in two ways—framework-to-framework and item-to-framework.

Framework-to-Framework Activities. The framework-to-framework activities involved matching the NAEP framework topics and subtopics for grade 8 within the five content strands to the seven North Carolina competency goals and objectives for grade 8. [See Kenney, et. al. (1998) for a complete discussion of these activities and the associated materials.] The six panelists

were divided into two groups (for replication purposes) to complete the activity and then were brought together for discussion and consensus. The independent agreement between the two groups was very high. Any major disagreements were adjudicated during the discussion phase and it was relatively easy for the groups to resolve the disagreements.

In general, the panelists agreed that there was moderate congruence with respect to the content characteristics of the two assessments based on the frameworks. Table 3 shows the results of the congruence between the competency goals and objectives of the North Carolina framework and the topics and subtopics of the NAEP framework.

Table 3. Level of Congruence Between Frameworks.

NAEP Content Strand	NC Competency Goal	Level of Congruence
Data Analysis, Statistics, & Probability	Probability and Statistics (Goal 6)	High
Measurement	Measurement (Goal 4)	Moderate
Algebra & Functions	Pre-Algebra (Goal 3)	Moderate
Geometry & Spatial Sense	Geometry (Goal 2)	Low
Number Sense, Properties, & Operations	Numeration (Goal 1) Computation (Goal 7)	Low

Item-to-Framework Activities. The item-to-framework activities involved the matching of items from one assessment to the framework of the other assessment. [See Kenney, et. al. (1998) for a complete discussion of these activities and the associated materials.] Panelists were asked to classify a set of NAEP items according to the North Carolina competency goals and objectives and to classify a subset of North Carolina items according to the NAEP topics and subtopics.

When classifying the NCEOG items to the NAEP framework, the panelists stated that the majority of the items could be classified into one or more NAEP topics. The panelists also noted that a few topics within a content strand were used repeatedly. For example, most of the NCEOG measurement items were classified into the NAEP topics of perimeter, area, volume, and surface area. The panelists concluded that the NCEOG grade 8 competency goals and objectives were a subset of the NAEP content strands and topics.

When classifying the NAEP items into the NCEOG framework, the panelists stated that a number of the items could not be classified. For example, a NAEP item about factors and multiples could not be matched to the NCEOG competency goals and objectives. Factors and multiples are found in the grade 5 NCEOG framework (Grade 5, Objective 1.3: Find multiples and factors of a number, explain the process.) and would be assessed on the grade 5 test in North Carolina.

Conclusion. Based on the content congruence activities, 18 of the 34 of the NAEP content topics for grade 8 (53%) matched with particular North Carolina competency objectives. The Data Analysis, Statistics, and Probability strand had the highest percent of topic matches (78%)

and the Geometry strand had the lowest percent of topic matches (38%).

The panelists did observe some differences between the two curricular frameworks. Ten of the 34 NAEP content topics (29%) and 7 of the 31 North Carolina competency objectives (23%) showed evidence of non-confirmation. Some of these differences could be explained based on the information provided by the panelists knowledge about each of the assessments. The two explanations for non-confirmation were: (1) the generality of the descriptions of the topics/objectives in the frameworks and (2) the generality of one framework description and the specificity of the other framework description.

The panel concluded that there was a moderate degree of congruence between the tests (average rating of 3 on a 5-point scale)—there are differences between the North Carolina test and the NAEP test at grade 8 along the content dimension. However, these differences were not sufficient to account for the magnitude of difference between proficient performance on the North Carolina test and proficient performance on NAEP.

Cognitive Dimension

The fourth phase of the study was designed to compare the cognitive characteristics of the NCEOG and NAEP assessments. In this study, the cognitive characteristics of a test refered to the extent to which the test engages students in various cognitive processes such as execution of procedures, recall of facts, conceptual understanding, and problem solving. The relationship between the two assessments was investigated in two ways—NAEP ability categories and cognitive demand.

The subset of NCEOG items used with these activities were chosen on the basis of the results from the content congruence activities. The content areas of probability and statistics, measurement, and pre-algebra were designated as areas for the analysis along the cognitive dimension. Because of the congruence between the NCEOG and NAEP in these content areas, it was important to investigate the degree to which the relationship between content areas extended to the cognitive characteristics of each test. The NCEOG set of items consisted of 59 of the 80 items on the test form, and the NAEP set of items (N = 48) consisted of three blocks that contained the greatest number of items classified in the three target content areas.

NAEP Ability Categories Activity. This activity involved matching the set of NCEOG items to the NAEP ability categories—Conceptual Understanding, Procedural Knowledge, and Problem Solving. "Conceptual knowledge can be viewed as a measure of the student's knowing 'that' or 'about,' while procedural knowledge can be viewed as a student's knowing 'how.' These two abilities combined provide a base for the capability to recognize and understand a situation, to formulate a plan to confront the situation, to arrive at a solution to a problem the situation presents, and to reflect upon the solution. These latter stages can be thought of as facets of problem s\. lving" (The College Board, 1994, p. 39). [See Kenney, et. al. (1998) for a complete discussion of this activity and the associated materials.]

Table 4 shows the results of matching the NCEOG items to the NAEP ability categories. About half of the items (48%) were classified by a majority of the panelists as Procedural Knowledge, with the remaining items divided nearly evenly between the other two categories.

Table 4. Percentage Distribution of a Set of North Carolina Items Classified According to the NAEP Ability Categories.

	Conceptual Understanding	Procedural Knowledge	Problem Solving
Overall (N = 59 items)	**27**	**48**	**25**
Numeration (1) & Computation (7)	11	82	6
Pre-Algebra (3)	33	44	22
Geometry (2)	43	29	29
Measurement (4)	0	29	-71
Problem Solving (5)	22	44	33
Probability and Statistics (6)	75	25	0

From Table 4 it can be seen that there are notable differences when the results are examined at the competency goal level. The goals involving real number concepts (Goals 1 and 7) had the highest percentage of items classified as procedural knowledge; the goal involving probability and statistics (Goal 6) had the highest percentage of items classified as conceptual understanding; and the goal involving measurement (Goal 4) had the highest percentage of items classified as problem solving. Only one-third of the items in the problem-solving goal (Goal 5) were classified as problem solving.

Cognitive Demand Activity. This activity involved comparing the NCEOG and NAEP items to external criteria that represented various levels of cognitive demand. The criteria were obtained from a variety of sources including *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989) and other studies involving NAEP (e.g., Romberg, Smith, Smith, & Wilson, 1992). The final set of criteria included those that represented both high levels of cognitive demand (problem solving, reasoning, communication, and connections) and those that represented low levels of cognitive demand (recall of facts, routine procedures, and estimation). [See Kenney, et. al. (1998) for a complete discussion of this activity and the associated materials.]

Table 5 shows the results of the cognitive demand activities. There was not much of a difference between the percent of items with high and low cognitive demand between the two assessments.

Table 5. Summary of Results from the Cognitive Demand Activity.

	High Cognitive Demand		Low Cognitive Demand	
	NCEOG	NAEP	NCEOG	NAEP
Overall	**48**	**52**	**52**	**48**
Number	13	57	87	43
Measurement	78	20	22	80
Geometry	67	29	33	71
Data Analysis, Statistics, & Probability	70	78	30	22
Algebra	42	77	58	23

From Table 5 it can be seen that there are notable differences when the results are examined at the competency goal level. For the three target content strands that the panelists identified as being highly or moderately congruent on the content dimension, only the Data Analysis, Statistics, and Probability strand was similar between the two assessments with respect to the cognitive dimension. Based on the panelist's judgments, the NCEOG measurement items were more cognitively demanding than the NAEP items, and the reverse was true for the algebra items.

Conclusion. The panel concluded that there are differences between the North Carolina test and the NAEP test at grade 8 along the cognitive demand dimension. However, these differences were not sufficient to account for the magnitude of difference between proficient performance on the North Carolina test and proficient performance on NAEP.

Discussion

Content alignment, or "congruence" as it is described in this paper and the other associated papers, is a process that examines the degree to which expectations and assessments are in agreement. The results from such a process can be used to guide the reform (or need for reform) of an educational system to ensure that students are learning what they are expected to know and do.

A content alignment study can be framed in terms of a variety of perspectives. It can be conducted entirely by an external group, entirely by an internal group, or somewhere in between. It can be conducted in reference to another assessment, in reference to only itself, or somewhere in between. Finally, a content alignment study can be conducted concerning only an assessment, only a curricular framework, or somewhere in between. This study took the "in between" posture on each one of the frames of reference for the study; consequently, the results from this study have the potential to be useful to a very large audience—the test developers, the curriculum developers, the test users, the policy makers, and the public-at-large.

What was it like to participate?

Participating in a content alignment study can be stressful, rigorous, challenging, and insightful all at the same time. Some of the feelings have to do with the perspective each of the panelists brings to the project. The content alignment model developed in this study brings together panelists with a variety of backgrounds—individuals intimately familiar with each assessment being examined and individuals who do not have knowledge of either assessment but have considerable knowledge of the construct being assessed. By having such diversity in the panel members perspectives, the conclusions are based on knowledge. Panelists knowledgeable about one of the assessments could provide answers to questions and clarify the thinking about the specific assessment.

When the assessment that a panelist is familiar with is being examined, often feelings of stress, frustration, and defensiveness are exhibited. After all, the individuals familiar with the assessment developed the assessment to be the "best" that it could be. When the assessment that a panelist is not familiar with is being examined, often feelings of understanding and insight are exhibited. It is much easier to understand the effect of cognitive demand on content when you did not develop the curricular framework or the assessment items.

What was learned?

Better Understanding of NAEP. One of the important results of this study was a better understanding of the NAEP assessment—its technical characteristics, content character, and level of cognitive demand. In North Carolina, the NAEP is looked upon as an important indicator of student achievement. It is the one valid of measure of achievement that enables us to examine how the students of North Carolina are achieving compared to the rest of the United States. Other nationally standardized assessments can only provide comparisons to a norming group that was tested sometime in the past (maybe as much as 8 to 10 years ago). But, from this study it was concluded that there are differences between the North Carolina test and the NAEP test at grade 8, but the differences were not sufficient to account for the magnitude of the difference between proficient performance on the North Carolina test and proficient performance on NAEP. This conclusion also helps to put NAEP into perspective: while NAEP is looked upon as "an" important benchmark of student achievement, it may not be "the" benchmark of student achievement.

Better Understanding of the NCEOG. The most important result of this study was a clearer, more unbiased understanding of the NCEOG assessment—its technical characteristics, content character, and level of cognitive demand. Because of this thorough review of the relationship of the North Carolina curricular framework to the NCEOG and both the framework and the test to the NAEP, several differences between the NCEOG and the NAEP assessments emerged.

The first major difference between the assessments concerns the test specifications and the developmental nature of the frameworks. The NCEOG assessment consists of a unique assessment at each grade (3 through 8) that assesses only the concepts taught at that specific grade (there is no overlapping content). The NAEP assessment consists of three overlapping assessments (grades 4, 8, and 12) and, within a test, the concepts covered are from those taught across a span of grades (grade 8 test covers concepts from grades 5 through 8). This developmental difference in content specification explains the low level of content congruence in the framework-to-framework activity for number sense, properties, and operations (NAEP) versus numeration and computation (NCEOG). For example, on the NAEP test number properties such as odd and even numbers and factors and multiples are assessed, but these

are not assessed on the NCEOG grade 8 assessment because they are taught at grades 3 and 5 respectively. This difference also explains the lack of congruence for some NAEP items in the item-to-framework activity. For example, a grade 8 NAEP item on factors and multiples is assessed in North Carolina at grade 5 and a grade 8 NAEP item on unit conversion is assessed in North Carolina at grade 6.

The second major difference between the assessments concerns the curricular frameworks—the level of specificity versus generality. On the NAEP assessment, symmetry is described in a very general, open way (Topic 3: Identify the relationship (congruence, similarity) between a figure and its image under a transformation; Subtopic a: Use motion geometry (informal: lines of symmetry, flips, turns, and slides)). Whereas, in the North Carolina curriculum the objective used for instruction and assessment is very specific in the method to be used to solve symmetry problems (Objective 2.2: Solve problems related to similar figures using indirect measures to determine missing sides.). This difference in the level of specificity explains the low level of content congruence in the framework-to-framework activities for geometry.

Another major difference between the assessments concerns the cognitive demand of specific parts of the curricular framework. Overall, the NCEOG and the NAEP had about the same proportion of high cognitive demand items and low cognitive demand items (refer to Table 5). Very diverse results were obtained when the level of cognitive demand of the two assessments was compared at the competency goal/topic level. The panelists observed that some of the differences may have to do with the number of steps needed to arrive at an answer or the format (multiple-choice versus constructed response) of the item.

This alignment study also provided a better understanding related to the cognitive nature of the NCEOG tests. The North Carolina curricular frameworks were developed using the work of Robert J. Marzano and his colleagues on the *Dimensions of Thinking* (1988). This thinking skills framework consists of 21 core thinking skills that are organized into 7 broad skills. While each of the 7 skills is fairly easy to discuss, it was very hard to classify a specific test item into one of the skills. The panelists in the study discussed this area and concluded with the following question: Can you say that all students will use the same single "thinking skill" to solve a problem? The panelists also noted that the level of cognitive demand of items could be related to the specificity of the curriculum. In the North Carolina curricular framework Pascal's triangle and the Fibinocci sequence are specifically stated and they are expected to be a part of the instruction. Consequently, for students in North Carolina these items would most likely be taping low levels of cognitive demand. In another state where these sequences are not specifically stated to be a part of instruction, a student would need to reason the problem out and the item would most likely exhibit high cognitive demand.

How will the information be used?

In order for any alignment study to be worthwhile, the results must be practical, useful, and capable of being acted upon. The results of this study can be applied to three areas of curricular and test development: (1) content character—curricular framework revisions, (2) content character and cognitive demand—test specifications, and (3) cognitive demand—item development.

The ability to act upon any of the results from this study is dependent on the political nature of the testing program in North Carolina. North Carolina has a very "high-stakes" testing program; decisions based on assessment data affect not only schools and school districts, but also students in terms of promotion and retention and teachers in terms of money and staff development. North Carolina has implemented the ABCs Accountability Program to reward

schools that are making progress and to provide targeted-assistance to those who are not. A part of this program has also led to the possibility of teacher-testing in schools that are not making progress. Many individuals want the North Carolina tests to be more "NAEP-like"; but what does this mean? Does this mean that the North Carolina End-of-Grade Tests should produce results that are similar to the results reported from NAEP? Does this mean that the test content and the tests themselves should look like NAEP? At the present time North Carolina has a curricular framework for each grade and the associated test; NAEP has a curricular framework that covers multiple grades. At the preset time North Carolina generally uses multiple-choice assessments for accountability purposes because it is desired that the tests be administered at the end of the instructional period (end of the school year), but also be reported immediately. Whereas, NAEP is administered in February and the results are typically reported a year later. With curricular changes currently underway, North Carolina has a window-of-opportunity to make changes in some parts of the program and to improve other parts.

Curricular Framework Revisions. In North Carolina the curricular frameworks are scheduled to be revised about every 5 years. The results from this study can be used to examine the "grain size" of the competency goals and objectives of the North Carolina *Standard Course of Study*. This study revealed that some of the competency objectives are very general (very large "grains" of content) and encompass a broad range of topics and levels of cognitive demand (for example, Objective 4.1: Estimate the answer; then solve complex problems that include application of measurement; determine precision and check for reasonableness of results.). Other competency objectives are very specific (very small "grains" of content) that will likely lead to specific instruction and less cognitive demand and problem solving (for example, Objective 4.5: Explore the effect on plane and solid figures when a dimension of a figure is changed.). These differences in "grain size" were accommodated in the test specifications by having more items assess the broader objectives and less items for the more specific objectives (Objective 4.1 was assessed by 3 items and Objective 4.5 was assessed by only 1 item).

Test Specifications. Users of an assessment rarely see more than one grade level or form of a test. They see the test that they will be administering that day; they see the specific test their students took; or they see the test that will be used to set performance standards for promotion. We as developers and curriculum specialists see the continuity of the testing program across the grades. In North Carolina, we don't expect to see grade 5 and 6 competency objectives tested on the grade 8 test. Conducting a content alignment study with only one grade of the North Carolina End-of-Grade Test helped us to understand "our client's" perspective. We need to pay more attention to each test at each grade as a separate entity and examine its scope and breadth.

Item Development. The results of this study will be very useful when examining items to keep for future use after a curriculum revision. Each item can be examined by a series of questions:

- What level of cognitive demand does each competency objective call for?
- If the objective specifies a level of cognitive demand, do all of the items match that level?
- If the objective does not specify a level of cognitive demand, does the range of items exhibit both high and low levels of cognitive demand?

In addition, the distractors for each multiple-choice item should be examined to determine if the cognitive demand of an item is maintained in the distractors. The distractors can actually reduce or increase the cognitive demand of an item depending on how well-chosen they are. From the results of the review of current items, specifications for further item development can be refined.

When the panelists categorized items as to level of cognitive demand, they noted two possible confounding factors—the number of steps in the problem and the format of the item. These two factors need to be examined further. During item review some information could be collected showing the number of steps required to solve the problem, the actually difficulty of the item, and the level of cognitive demand. If it is found that the number of steps is related to the cognitive demand, then items could be developed that are more novel and require more actual "problem solving skills," but, at the same time, only require one or two steps to solve.

Further Areas of Research

The main conclusion of this study was that while there are differences between the NAEP and NCEOG, these differences are not sufficient to explain the magnitude of differences in performance on the two assessments. Two areas of further research are needed to better understand these differences in performance:

1. Investigate the performance standards associated with each assessment—the definitions of the standards, the method used to set the standards (task-centered vs. examinee-centered), and the consequences of the standards.
2. Investigate the effect of the computation section on the NCEOG compared to the extended constructed-response items on the NAEP. Does the use of extended constructed-response items make a difference?

The first area of research concerns the performance standards associated with each assessment. The wording of each of the descriptors of the standards will have an impact on the rest of the standard-setting process. For NAEP, the general policy definition for the proficient achievement level includes "competency over challenging subject matter" (Reese, Miller, Mazzeo, & Dossey, 1997), whereas, in North Carolina the definition of proficient refers to grade-level knowledge and skills. The standard-setting methods employed will also have an impact on the final standards developed. Task-centered methods tend to set standards at the ends of the distribution either too high or too low. Examinee-centered methods are based on what examinees "are able to do" rather than what they "should be able to do," and may led to standards being set somewhat lower overall.

Under the North Carolina Accountability Program, the North Carolina End-of-Grade Tests are high-stakes assessments for teachers, schools, and school districts. The grade 8 assessments also have high-stakes consequences for students. The grade 8 test serves as a screening for taking the competency test in high school. To receive a North Carolina High School diploma, all students must achieve Level III on the grade 8 assessment or the North Carolina Competency Test. In comparison, the NAEP assessment is perceived as a low-stakes assessment by students, teachers, schools, and school districts. The one group that does not perceive the NAEP as being low stakes is state departments of education because of perceptions and beliefs held by the public and policy-makers. The level of consequences for the actual test-takers (the students) will have some impact on the level of performance when taking each assessment. Based on field test results where students understand that the test is a low-stakes assessment, there is typically a 5- to 10-point increase in scores compared to when the assessments are perceived as high-stakes (during the actual statewide administration).

The second area of investigation is to examine the places where the NCEOG and NAEP are different—the NCEOG computation section and the NAEP extended-constructed response items. While both of these sections of the tests are small, they should be investigated to see what impact they have on the overall performance of students.

References

Archer, J. (1997, January 15). "States struggle to ensure data make the grade." *Education Week on the WEB, 16.*

The College Board. (1994). *Mathematics framework for the 1996 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board.

Kenney, P.A., Silver, E.A., Alacaci, C, & Zawojewski, J.S. (1998). *Content analysis project—State and NAEP mathematics assessments.* Washington, DC: National Assessment Governing Board.

Kenney, P.A. & Silver, E.A. (1999, April). Design features for the content analysis of a state assessment and NAEP. Paper presented at the Annual Meeting of the American Educational Research Association.

Marzano, R.J., Brandt, R.S., Hughes, C.S., Jones, B.F., Presseisen, B.Z., Stuart, C., & Suhor, C. (1988). *Dimensions of Thinking.* Alexandria, VA: Association for Supervision and Curriculum Development.

Musick, M.D. (1996, June). *Setting educational standards high enough.* Atlanta: Southern Regional Education Board.

National Assessment Governing Board. (1996, August). *Policy statement on redesigning the National Assessment of Educational Progress.* Washington, DC: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

North Carolina Department of Public Instruction. (1989). *North Carolina Standard Course of Study for Mathematics.* Raleigh, NC: Author.

North Carolina Department of Public Instruction. (1996). *North Carolina End-of-Grade Tests, Technical Report #1.* Raleigh, NC: Author.

North Carolina Department of Public Instruction. (1997, February 12). North Carolina Assessments and NAEP: Linking issues and options. Minutes from NCDPI and NAGB meeting in Raleigh, NC.

North Carolina Department of Public Instruction. (1999, March 4). "Reading Focus Delivers Results." *North Carolina Public Schools InFoWeb.* Raleigh, NC: Author.

Reese, C.M., Miller, K.E., Mazzeo, J. & Dossey, J.A. (1997, February). *NAEP 1996 mathematics report card for the nation and the states: Findings from the National Assessment of Educational Progress.* Washington, DC: National Center for Educational Statistics.

Romberg, T.A., Smith, M., Smith, S., & Wilson, L. (1992, June). The feasibility of using international data to set achievement levels for the National Assessment of Educational Progress (NAEP). Madison, WI: National Center for Research in Mathematics Education.